DOCUMENT RESUME

ED 175 904 TM 009 260

AUTHOR Yap, Kim Onn; And Others
TITLE Effects of Data Analysis Methods and Selection Procedures in Regression Models.
PUB DATE Apr 79
NOTE 46p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bias; Compensatory Education Programs; Evaluation Methods; *Mathematical Models; Multiple Regression Analysis; *Program Evaluation; *Research Design; *Research Problems; *Sampling
IDENTIFIERS Elementary Secondary Education Act Title I; *RMC Models

ABSTRACT

The effects of using different data analysis methods on estimates of treatment effects of educational programs were investigated. Various regression models, such as those recommended for Title I program evaluations, were studied. The first effect studied was the amount of bias that might be expected to occur in the various settings. Results indicated that bias was rarely a problem. What did appear to be a problem was the efficiency of the estimates. That is to say that while the estimates were not consistently off of the expected true value (bias), there was considerable inaccuracy in the results (inefficiency). Several recommendations were made with respect to selecting and placing into treatment and control groups. Furthermore, some concern was expressed over the use of the regression model to evaluate educational programs at all. (JKS)

EFFECTS OF DATA ANALYSIS METHODS AND SELECTION PROCEDURES IN REGRESSION MODELS

Kim Onn Yap

Gary D. Estes

Joe B. Hansen

Northwest Regional Educational Laboratory

A paper presented at the Annual Meeting of the American Educational Research Associati .., San Francisco, April 8-12, 1979

EFFECTS OF DATA ANALYSIS METHODS AND SELECTION PROCEDURES IN REGRESSION MODELS

Kim Onn Yap

Gary D. Estes

Joe B. Hansen

Northwest Regional Educational Laboratory

INTRODUCTION

The use of experimental and quasi-experimental designs in educational program evaluations has resulted in numerous problems. An example of a large-scale effort to develop an evaluation system based on such design concepts is the attempt by the U.S. Office of Education to implement the Title I Evaluation and Reporting System (TIERS) on a national basis. TIERS was developed for the purpose of providing comparable data on the impact of Title I programs across projects. As described by Tallmadge and Horst (1976), Horst, Tallmadge and Wood (1975) and Tallmadge and Wood (1976, 1978), the system consists of (a) the norm-referenced models, (b) the control group models and (c) the regression models. The control group and regression models are essentially variations of designs described by Campbell and Stanley (1963) and Sween (1971).

The models are proposed for use in evaluating public school programs. However, it is not surprising that problems arise when these models are used in a loosely controlled educational setting. Educational

programs are often not structured in ways readily amenable to experimental design constraints.

According to Campbell and Stanley (1963), the regression models are most applicable when selection is made on the basis of a cutting score on a quantified composite of qualifications. The standard procedure in Title I programs is to select the most needy students for participation. Most frequently, students with the lowest scores on an achievement test, teacher ratings or some composite of similar data are selected to receive Title I services.

An explicit criterion for implementing the regression models correctly is that a single cut-off score be used to select students into the program, i.e., all students below the cut-off score are program participants and no students above the cut-off score are selected. Additionally, Tallmadge and Wood (1976, 1978) recommend that there be a reasonably high correlation between the selection measure and the criteria for program evaluation. More specifically, they recommend as a minimum a pretest-posttest correlation of at least .40 in the comparison group. It is also recommended that reasonably large sample sizes be used to ensure accurate estimates of treatment effects.

The use of regression models in program evaluations has attracted the attention of a number of investigators. Mandeville (1978) and Echternacht (1978), for example, compared results obtained from the norm-referenced and regression models and found that the results were not comparable. In his investigation of the use of true vs. observed pretest scores for selection, Goldberger (1972) demonstrated that selection based on observed pretest scores provided unbiased estimates of treatment

effects. Estes and Anderson (1978) studied a number of applications of the regression design in program evaluations and found that the design was sensitive to floor or ceiling effects.

The basic problem in implementing the regression models in program evaluations is that a strict cut-off score is often not used in selecting program participants. Several situations arise when school districts attempt to use the regression design, including:

Case 1: This is the ideal case in which a strict cut-off score is used for selection. Students scoring below the cut-off score are assigned to the Title I or treatment group and students scoring above the cut-off score do not receive treatment and serve as comparison students.

Case 2 A cut-off band occurs rather than a strict cut-off. The band is a result of students within a range around the cut-off score being randomly assigned to the treatment or comparison group. This occurs when a cut-off score is identified, but factors such as scheduling problems and unequal numbers of students across classrooms or schools result in a situation where some students below the cut-off score do not receive Title I assistance and some students above the cut-off score receive Title I assistance. This variation is characterized as random in that students in the cut-off band or fuzzy cut-off are not placed in treatment or comparison groups on any systematic or measured basis.

Case 3: Often students below the cut-off score are denied Title I services and students above the cut-off score are given Title I services on the basis of another variable, e.g., teacher ratings or judgments. This case differs from the second in that a systematic judgment or measured variable is used in creating a cut-off band or fuzzy cut-off.

Case 1 is the standard regression design, and data analysis procedures have been provided by Tallmadge and Wood (1976, 1978). There are, however, at least two ways to handle data obtained in Cases 2 and 3:

1. Students who fall within the cut-off band are excluded from the analysis.
2. Students who fall within the cut-off band are included in the analysis. They are treated as treatment or comparison group students as they had been assigned.

The above conditions give rise to four data analysis situations, namely:

1. A strict cut-off is used to assign students to Title I and comparison groups, and procedures outlined by Tallmadge and Wood (1976) are used to conduct data analysis.
2. There is a fuzzy cut-off, and students in the cut-off band are summarily excluded from data analysis.
3. There is a fuzzy cut-off, students in the cut-off band being assigned randomly to Title I and comparison groups. All students are included in data analysis and treated as Title I or comparison students as they had been assigned.

4. There is a fuzzy cut-off, students in the cut-off band being assigned to Title I and comparison groups on the basis of teacher ratings. All students are included in data analysis and treated as Title I or comparison students as they had been assigned.

The simulation study reported in the remainder of this paper was designed to assess the effects of these data analysis situations on estimates of treatment effects obtained with the use of the regression models.

PROCEDURES

Constructing the Variables

To study the effects of data analysis situations on estimates of treatment effects, data resembling those suited for analysis with the regression models were simulated. The rudiments of the simulation were as follows:

$$Y_{1ij} = X_{ij} + E_{ij}, \qquad (1)$$

$$Y_{2ij} = X_{ij} + G_{ij} + TE_{ij} + E'_{ij}, \qquad (2)$$

$$Z_{ij} = X_{ij} + E''_{ij}, \qquad (3)$$

where Y_{1ij} is the pretest score of student i in group j; Y_{2ij} is the posttest score of student i in group j; Z_{ij} is a teacher rating score for student i in group j; X_{ij} is the true achievement level of student i in group j at pretest; G_{ij} is the growth attributable to factors other than the treatment for student i in group j; TE_{ij} is the treatment effect for student i in group j; and E_{ij}, E'_{ij} and E''_{ij} are error terms.

For purposes of the simulation, it was assumed that the mean growth rates (G_{ij}'s) for the treatment and comparison groups are equal. In equation (2), TE_{ij}'s were set to equal zero for students in the comparison group to indicate the absence of treatment effects.

The values of X_{ij}, G_{ij}, TE_{ij}, E_{ij}, E'_{ij} and E''_{ij} were made up of random numbers provided by GAUSS (IBM, 1968), a computer subroutine which generates normally distributed random numbers. The relative size of X_{ij}, E_{ij}, E'_{ij} and E''_{ij} were adjusted by means of multipliers. For example, the values of a set of X_{ij}, E_{ij}, E'_{ij} and E''_{ij} may be obtained as follows:

$X_{ij} = .7\ N_1,$

$E_{ij} = .3\ N_2,$

$E'_{ij} = .3\ N_3,$

$E''_{ij} = .3\ N_4$, where

the Ns are random numbers. Means and standard deviations for the random numbers were chosen in such a way that Y_{1ij}, Y_{2ij} and Z_{ij} would have approximately a mean of 50 and a standard deviation of 21.06, respectively, to correspond with the mean and standard deviation of Normal Curve Equivalents (NCEs). For example, in

$Y_{1ij} = X_{ij} + E_{ij}$, where

$X_{ij} = .7\ N_1$, and

$E_{ij} = .3\ N_2,$

both N_1's and N_2's were given a mean of 50 and a standard deviation of 27.65. This gave Y_{1ij} a mean of 50 and a standard deviation of 21.06, i.e., $21.06 = \sqrt{(.7)^2\ (27.65)^2 + (.3)^2\ (27.65)^2}$. The same procedure was used to give Y_{2ij} and Z_{ij} a mean of 50 and a standard deviation of 21.06.

Means and standard deviations for G_{ij} and TE_{ij} were determined by providing the appropriate parameter values to subroutine GAUSS. G_{ij} was set to have a mean of 10 and a standard deviation of 10 and TE_{ij} was set to have a mean of 7 and a standard deviation of 7. These means and standard deviations had been chosen to reflect what is most likely to occur in real-life situations in terms of NCE scores.

Negative values provided by GAUSS, which occurred on few occasions, were dropped, resulting in slightly higher means and lower standard deviations for the variables.

Data Characteristics

Three parameters relating to data characteristics were manipulated in the simulation (see Table 1). First, data reliability was varied from .84 to .69. Second, the size of correlation between pretest (Y_{1ij}) and teacher ratings (Z_{ij}) was varied from .75 to .50. Third, sample size was made to vary from 100 to 200.

The manipulation of data reliability was based on Gulliksen's (1950) idea that a reliability coefficient can be expressed as the ratio of true variance to total variance. This means that we could vary reliability by applying different multipliers to the random numbers which make up the values of variables. For example, in

$$Y_{1ij} = X_{ij} + E_{ij}, \text{ where}$$

$$X_{ij} = .7\, N_1,$$

$$E_{ij} = .3\, N_2,$$

and N_1's and N_2's are given the same variance, the reliability coefficient of Y_{1ij} is given by

$$r_{Y_1 Y_1} = \frac{\text{Var } X_{ij}}{\text{Var } X_{ij} + \text{Var } E_{ij}} .$$

Since multiplying a set of numbers by a constant increases the variance by the square of the constant and since N_1's and N_2's have the same variance, we have

$$r_{Y_1 Y_1} = \frac{(.7)^2}{(.7)^2 + (.3)^2} = .84$$

That is, the reliability of Y_{1ij} is .84.

It could readily be verified that by changing the multipliers to .6 for N_1's and .4 for N_2's we will have lowered data reliability to .69. In the simulation, data sets with reliability (for both pretest and posttest) of .69 and .84 were created.

Correlations between pretest (Y_{1ij}) and teacher ratings (Z_{ij}) were controlled by means of the following formula:

$$R_{\infty\infty} = \frac{r_{1I}}{\sqrt{r_{11}\ r_{II}}}$$

Reported by Gulliksen (1950, p. 101), the formula gives the correlation between a test and a criterion when each is increased to infinite length to attain a reliability of unity. Given that

$$Y_{1ij} = X_{ij} + E_{ij}, \text{ and}$$

$$Z_{ij} = X_{ij} + E'_{ij},$$

the two variables share a single true score component with $R_{\infty\infty}$ reaching unity when both Y_{1ij} and Z_{ij} are made perfectly reliable. It follows that $\sqrt{r_{11}\ r_{II}} = r_{1I}$, which provides a means of obtaining a desired value of r_{1I} by changing either r_{11}, r_{II} or both.

In the simulation, we have required that r_{11} (reliability of Y_{1ij}) be either .84 or .69 (a fixed value), leaving r_{II} (reliability of Z_{ij}) to be varied to yield a desired value for r_{1I}. The way in which a desired correlation between Y_{1ij} and Z_{ij}, say .75, was obtained is illustrated as follows:

Since (a) $\sqrt{r_{11}\ r_{II}} = r_{1I}$, (b) the desired value of r_{1I} was .75 and (c) r_{11} had been given a reliability of .84, we had

$$\sqrt{(.84)\ (r_{II})} = .75$$

$$\sqrt{r_{II}} = \frac{.75}{\sqrt{.84}}$$

$$= .82$$

$$r_{II} = .67$$

In other words, giving Z_{ij} a reliability of .67 produced a correlation of .75 between Y_{1ij} and Z_{ij}.

Since the variance of Z_{ij} was made to equal 443.52 (the square of 21.06) the true variance required to yield a reliability coefficient of.67 was (443.52) (.67) which equals 297.16. An appropriate multiplier (.62 in this case) was then applied to X_{ij} in Z_{ij} (X_{ij} had a standard deviation of 27.65 when Y_{1ij} and Y_{2ij} were given a reliability of .84) to produce the required true variance.

Cut-off Location and Width

Two parameters relating to the selection and proportions of treatment and comparison groups were manipulated in the simulation. First, the width of the cut-off band was varied. When there is a strict cut-off, the width is zero. As more cases fall within the cut-off band, its width becomes greater. Two widths were used in the study: data sets with 10 percent and 20 percent of the simulated cases falling within the cut-off bands were created.

The second parameter was the location of the cut-off. Unless the cut-off band is exceedingly wide, the location of the cut-off determines the proportions of students assigned to the treatment and comparison groups. In the study the location of cut-offs were varied from the 20th to the 30th percentile point.

Creating the Data

The pretest data (Y_{1ij}) were first simulated. The hypothetical cases in each data set were rank-ordered. Cut-offs of different widths described earlier were then used to assign students to the treatment ($j = 1$) amd comparison ($j = 2$) groups. In the case of fuzzy cut-offs (i.e., when the width of the cut-off band was non-zero), assignments were made either randomly or on the basis of teacher ratings. When random assignment was used, random numbers were drawn from a table of random numbers to assign cases within the cut-off band to treatment and comparison groups. When assignment was made on the basis of teacher ratings, cases within the cut-off band were rank-ordered according to teacher ratings and then assigned to treatment and comparison groups.

As indicated earlier, the cut-off bands varied from a width covering 10 percent of the cases to a width covering 20 percent of the cases in each data set. The mid-points of these cut-off bands were located at the 20th and 30th percentile points.

After the hypothetical cases had been assigned to treatment or comparison groups, posttest data (Y_{2ij}) were simulated by means of equation (2), adding growth (G_{ij}) and treatment effects (TE_{ij}) to pretest scores of students receiving treatment and only growth (G_{ij}) to pretest scores of comparison students.

The use of a variable rather than a constant as treatment effects was done to simulate what is most likely to occur in real-life situations. Treatment effects in real life undoubtedly vary from individual to individual within the treatment group. While the use of a variable will not produce results different from what one would obtain with the use of a constant, the use of a variable seemed conceptually more satisfactory.

THE DATA SETS

To study the impact of the various parameters (see Table 1) on the estimation of treatment effects, a variety of data sets were created. Taking into account the different levels of each of the three parameters relating to data characteristics (i.e., data reliability, size of correlation between pretest and teaching ratings, sample size), a total of 8, i.e., 2 x 2 x 2, categories of data sets were simulated. One hundred data sets were created for each of the categories. Characteristics of these data sets are summarized in Appendices A to H.

Table 1 about here

Since we had two cut-off points (at 20th and 30th percentiles) and two widths for the cut-off bands (10 and 20 percent of cases), each category of data sets in effect provided four different groupings of treatment and comparison students. Thus, a total of 32, i.e., 8 x 4, data classifications, each replicated 100 times, were simulated in the study.

Characteristics of the simulated data suggests that they closely resembled what we had intended to create. The obtained values, in some instances, deviated slightly from the parameters. As explained earlier, this came about essentially as a result of dropping negative values provided by GAUSS on a few occasions. Except for the slightly higher means and lower standard deviations, the data have the appearance of NCE scores. (The higher means of Y_{2ij} are due to higher means for X_{ij} and G_{ij}.) In summary, the observed characteristics of the data sets provided evidence that subroutine GAUSS and subsequent manipulation produced the desired data.

ANALYSIS AND RESULTS

For each of the 32 data classifications, the four data analysis situations described earlier were simulated:

1. Strict cut-off. There was a strict cut-off at the 20th or 30th percentile point. Procedures described by Tallmadge and Wood (1976, 1978) were used to analyze the data.
2. Leave-out. There was a fuzzy cut-off and cases in the cut-off band were excluded from data analysis.
3. Random selection. There was a fuzzy cut-off, and cases in the cut-off band were assigned randomly to Title I and comparison groups. All cases were included in data analysis and were treated as Title I or comparison students as they had been assigned.
4. Teacher selection. There was a fuzzy cut-off, and cases in the cut-off band were assigned to Title I and comparison groups on the basis of teacher ratings. All cases were included in data analysis and were treated as Title I or comparison students as they had been assigned.

In each of the data analysis situations a regression line was determined on the basis of comparison group data in order to predict what the performance of the Title I group would have been if there had been no Title I treatment. The prediction was made at the point where the treatment group's pretest mean intercepted the regression line. The predicted performance was then subtracted from the actual performance of the treatment group with the remainder being the estimated treatment effect or gain.

The estimated gain was then subtracted from the actual gain (TE_{ij}) which was built into the posttest (Y_{2ij}) of the treatment group. The difference was interpreted as an index of the accuracy with which the regression models estimate treatment effects in each of the four data analysis situations. The means and standard deviations of such differences by data categories by data analysis situations and by data classifications are summarized in Tables 2-9.

<u>Tables 2-9 about here</u>

DISCUSSION

Before we examine the effects which data analysis situations and the manipulated parameters have on the estimation of treatment effects, it might be helpful to present a perspective in which the results will be interpreted. As Wonnacott and Wonnacott (1970) point out, an estimator can be described in terms of bias, efficiency and consistency. An unbiased estimator is one that is, on the average, right on target. In other words, its expected value is identical with the true value of the parameter. A biased estimator, on the other hand, has an expected value that is "off target" or deviates from the true value of the parameter. An efficient estimator is an unbiased estimator with a relatively small variance. An inefficient estimator, on the other hand, is an unbiased estimator with a relatively large variance. A consistent estimator is one which zeroes in on the true value of the parameter as sample size increases.

Bias

Viewed in this perspective, the results in Tables 2-9 are evidence that the regression models provide relatively unbiased estimates of treatment effects when a strict cut-off was used for selection. The mean differences between the estimated and actual gains were in general negligibly small. Only in two instances (in Category V data sets) did the mean difference exceed an absolute value of 1.0. While the estimates could be considered to be practically unbiased in all cases, a shift of the cut-off from the 20th to the 30th percentile point appeared to further reduce the already small amount of bias. An increase in the

total sample size from 100 to 200 did not seem to have any appreciable or systematic effects on the amount of bias in estimation. The same also appeared to be true of an increase in data reliability from .69 to .84.

When the width of the cut-off band was non-zero (i.e., when a fuzzy cut-off was used), excluding all cases in the cut-off band or fuzzy area appeared to be a reasonable procedure to follow. In most cases, the difference between estimated and actual gains was shown to be less than an absolute value of 1.00. In no instance did the difference reach an absolute value of 2.00, the highest value being -1.58 (see Table 6).

There was a slight tendency for the difference between estimated and actual gains to decrease when the cut-off was moved from the 20th to the 30th percentile point. The width of the cut-off band did not seem to have any systematic or appreciable effects on the amount of bias. The same was true of an increase in data reliability from .69 to .84. Increasing the total sample size from 100 to 200 did not produce any appreciable differences in the amount of bias in estimation.

In the third analysis situation where cases in the cut-off band were randomly assigned to treatment and comparison groups, practically no bias was introduced in the estimation of treatment effects. In no instance was the difference between estimated and actual gains greater than an absolute value of 1.0. In this data analysis situation, neither the location of the cut-off nor the width of the cut-off band had any appreciable or systematic effects on bias. This was also true of an increase in total sample size from 100 to 200 and an increase in data reliability from .69 to .84.

In the fourth data analysis situation where cases in the cut-off band were assigned to treatment and comparison groups on the basis of teacher ratings, the results were a little different. The amount of bias, to begin with, was more substantial than that found in the first three data analysis situations. As a matter of fact, in almost half of the instances, the difference between estimated and actual gains was found to be greater than 1.0 in absolute value. In a few cases, the difference exceeded 2.0, the greatest difference being 3.36 (see Table 4). Both the location of the cut-off and the width of the cut-off band were shown to have considerable effects on the amount of bias in estimation. Bias was shown to increase when the cut-off was moved from the 20th to the 30th percentile point or when the width of the cut-off band was increased from 10 percent to 20 percent of the total sample. Differences produced by an increase in the width of the cut-off band were quite conspicuous.

In this data analysis situation, data reliability was shown to have a bearing on bias. As would be expected, less bias was found in data sets with a higher level of reliability than in data sets with a lower level of reliability. The difference was quite substantial in some cases (e.g., 1.53 vs. 2.93 and 1.88 vs. 3.36 in Tables 2 and 4).

An unanticipated outcome was that there seemed to be an inverse relationship between the size of correlation between pretest and teacher ratings on the one hand and the amount of bias on the other. More specifically, an increase in correlation from .50 to .75 actually produced greater differences between estimated and actual gains. This was true across all data categories. An increase in the total sample size from 100 to 200 appeared to have negligible effects on bias.

If it were possible to make a summary statement on the amount of bias produced in the various data analysis situations, it would be that with the exception of situation four (where teacher ratings were used to assign students in the cut-off band) such bias, if it existed at all, tended to be negligibly small. In a predominant majority of the cases, the difference between estimated and actual gains was less than 1.0 in absolute value. Interestingly enough, where bias was found to exist, it generally favored the treatment group in that the estimated gain was higher than the actual gain. On the other hand, bias introduced by the use of teacher ratings (which was found to be quite substantial in some cases) generally suppressed treatment effects in yielding an estimated gain that was less than the actual gain.

Efficiency

Did the four data analysis situations provide estimates of treatment effects that were equally efficient? A close scrutiny of the results suggests that the answer is no. Systematic differences did exist in the standard deviations of the mean differences between estimated and actual gains.

The results showed that, overall, the smallest standard deviations were found in the random selection situation, making its estimates the most efficient of the four data analysis situations. The largest standard deviations were found in the leave-out situation (where cases in the cut-off band were excluded from data analysis), making its estimates the least efficient. Estimates obtained in the strict cut-off and teacher selection situations appeared to be highly similar in terms of efficiency. The relative efficiency of estimates obtained in the four

data analysis situations appeared to hold across all eight data categories.

Of the parameters manipulated in the study, it appeared that data reliability and sample size undoubtedly had some effect on the efficiency of the estimates. Lower reliability and smaller sample size generally resulted in a decrease in efficiency, with the effects of sample size being a little more conspicuous than that of data reliability. Neither the location of the cut-off nor the width of the cut-off band was shown to have any systematic or appreciable effects on the efficiency of the estimates. Similar findings were obtained for all four data analysis situations. In the teacher selection situation, an increase in the correlation between pretest and teacher ratings from .50 to .75 did not seem to have made any appreciable difference in terms of efficiency of the estimates.

Perhaps a significant and somewhat unanticipated finding was that the random selection situation was shown to have provided estimates that were as efficient as (if not more so than) those obtained in the strict cut-off situation. Furthermore, the estimates obtained through the use of a strict cut-off were not as efficient as one would have expected. This was particularly true when the sample size was small, say 100. A standard deviation of 3 to 5 points produces a confidence interval of 12 to 20 points at about the .05 level. Confidence intervals of that magnitude can hardly be depended upon to accurately assess small gains in a single evaluation of compensatory education programs.

Consistency

As indicated earlier, an increase in sample size from 100 to 200 was found to enhance the efficiency of the estimates quite considerably.

However, there was no evidence that the increased sample size also reduced the amount of bias at the same time. In fact, in most cases, the reverse was found to be true. That is, there appeared to be a slight increase in bias with the larger sample size. Thus, when the regression models are used to assess project impact none of the four data analysis situations simulated in the study would provide consistent estimates of treatment effects.

SUMMARY AND CONCLUSIONS

The primary purpose of the simulation was to compare estimates of treatment effects in four different data analysis situations. We first examined the amount of bias that can be expected to occur in each data analysis situation. As it turned out, bias was not shown to be a major problem. With the exception of situations where teacher ratings were used as the basis for assigning students in the cut-off band to treatment and comparison groups, the amount of bias, if it existed at all, was shown to be negligibly small. Even in the teacher selection situation the amount of bias was in most cases of little practical import.

Most of the parameters manipulated in the simulation did not seem to have any systematic effects on the amount of bias. A notable exception was the width of the cut-off band. A greater width seemed to introduce a greater amount of bias, as would be expected. That is, the larger the fuzzy area, the further off was the estimate from the target value.

What appeared to be a real problem was the efficiency of the estimates. The standard deviations of mean differences between estimated and actual gains across all four data analysis situations ranged from slightly more than 2 to slightly less than 7. At the .05 level of significance this range covers confidence intervals of 8 to 28 points. Intervals of such magnitude clearly cannot be depended upon to provide an accurate assessment of small ach vement gains typically made by Title I students.

An unanticipated finding with respect to efficiency was that estimates obtained in the strict cut-off situation were not necessarily

more efficient than estimates obtained in the other situations. Standard deviations of mean differences obtained in the strict cut-off situation ranged approximately from 3 to 5 points, providing confidence intervals of 12 to 20 points at the .05 level. Needless to say, such intervals would appear to be too wide for assessing achievement gains in a local program evaluation with 100-200 students.

The least efficient estimates were obtained when cases in the cut-off band were excluded from data analysis. It should be noted, however, that when sample size was increased to 200, estimates obtained in the leave-out situation were found to be as efficient as those obtained in the straight cut-off situation with a sample size of 100.

In all data analysis situations sample size was found to be the major contributing factor to increased efficiency. Standard deviations of mean differences between estimated and actual gains decreased quite considerably (generally from 1 to 2 points) when sample size was increased from 100 to 200. There was, however, no evidence that estimates provided by the regression models were consistent estimates. In other words, an increased sample size did not seem to render the amount of bias smaller.

These findings make it rather difficult to formulate a hard and fast guideline for using fuzzy cut-offs. However, the results do appear to support a few rules of thumb:

1. If assignment of students in the cut-off band is random, estimates of treatment effects may be obtained by including all students in data analysis and treating the students as Title I or comparison students as they had been assigned. Estimates

obtained in this manner will not be any more biased and will probably be more efficient than estimates provided under the strict cut-off situation.

2. When students in the cut-off band are assigned to treatment and comparison groups on the basis of a third variable, say, teacher ratings, it would appear reasonable to estimate treatment effects by excluding students in the cut-off band from data analysis. This appears to be a reasonable rule when the cut-off band is relatively small (e.g., when it covers less than 10 percent of the treatment and comparison students) and when the sample size is relatively large (e.g., N = 200). In doing so, the evaluator can generally expect to come up with estimates which are not severely biased or less efficient than estimates obtained in other data analysis situations.

3. If students in the cut-off band are assigned to treatment and comparison groups on the basis of a third variable, say teacher ratings, and all students are included in the analysis, then estimates of treatment effects can be expected to be somewhat biased. This is particularly so when the cut-off band is relatively large (e.g., covering 20 percent of the treatment and comparison students).

4. Since none of the four data analysis situations can be expected to provide highly efficient and consistent estimates, evaluation results obtained through the use of the regression models must be interpreted with some degree of caution especially at the local level with relatively small sample sizes. The confidence

intervals for such estimates are quite large, considering the small amount of achievement gain typically produced by treatment in compensatory education programs. The results of the study suggest that when the regression models are used to estimate treatment effect, it would make sense to conduct significance tests, such as that described by Tallmadge and Horst (1976, p. 64), on the results before any conclusions are drawn with respect to treatment effects.

REFERENCES

Campbell, D. T., and Stanley, J. C. Experimental and quasi-experimental design for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Echternacht, G. J. The use of different models in the ESEA Title I evaluation system. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Estes, G. D. and Anderson, J. I. Observed treatment effects with special regression evaluation models in groups with no treatment. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Goldberger, A. S. Selection bias in evaluating treatment effects: Some formal illustrations. Institute for Research on Poverty. Discussion Papers, University of Wisconsin-Madison, April 1972.

Gulliksen, H. Theory of mental tests. New York: Wiley and Sons, 1950.

Horst, D. P., Tallmadge, G. K. and Wood, C. T. A practical guide to measuring project impact on student achievement. Mountain View, California: RMC Research Corporation, 1975.

IBM. System/360 scientific subroutine package. New York: IBM, 1968.

Mandeville, G. K. An evaluation of Title I Model C-1: The special regression model. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Sween, J. A. The experimental regression design--An inquiry into feasibility of non-random treatment allocation. Unpublished doctoral dissertation, Northwestern University, 1971.

Tallmadge, G. K. and Horst, D. P. A procedural guide for validating achievement gains in educational projects. Mountain View, California: RMC Research Corporation, 1976.

Tallmadge, G. K. and Wood, C. T. Users guide: ESEA Title I evaluation and reporting system. Mountain View, California: RMC Research Corporation, 1976.

Tallmadge, G. K. and Wood, C. T. Users guide: ESEA Title I evaluation and reporting system (Revised). Mountain View, California: RMC Research Corporation, 1978.

Wonnacott, R. J. and Wonnacott, T. H. Econometrics. New York: Wiley and Sons, 1970.

Table 1

Parameters Manipulated in the Simulation

Parameter	Level
1. Data reliability (r_{xx})	.69, .84
2. Correlation between pretest and teacher ratings ($r_{y,z}$)	.50, .75
3. Sample size (N)	100, 200
4. Width of cut-off band	10%, 20%
5. Location of cut-off point	20 %ile, 30 %ile

Table 2

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category I Data Sets ($r_{xx} = .84$, $r_{y,z} = .75$, $N = 100$)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.35	8.72	-.63	4.27
	20	10	9.75	9.12	-.63	4.27
	30	20	9.26	9.18	-.07	3.92
	30	10	9.23	9.16	-.07	3.92
Leave-out (Fuzzy cut-off)	20	20	9.91	8.86	-1.04	5.21
	20	10	9.84	9.19	-.65	4.68
	30	20	9.29	8.91	-.38	5.46
	30	20	9.59	9.16	-.43	4.64
Random Selection (Fuzzy cut-off)	20	20	9.54	8.92	-.62	3.97
	20	10	9.29	8.97	-.32	3.85
	30	20	9.16	9.01	-.15	4.13
	30	10	9.33	9.01	-.32	3.87
Teacher Selection (Fuzzy cut-off)	20	20	7.43	8.96	1.53	4.08
	20	10	8.44	9.02	.58	4.07
	30	20	7.18	9.06	1.88	3.76
	30	10	8.42	9.14	.72	3.78

Table 3

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category II Data Sets ($r_{xx} = .84$, $r_{y,z} = .50$, N = 100)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.24	8.72	-.52	4.26
	20	10	9.64	9.12	-.52	4.26
	30	20	8.94	9.18	-.24	4.13
	30	10	8.92	9.16	-.24	4.13
Leave-out (Fuzzy cut-off)	20	20	9.10	8.86	-.24	5.62
	20	10	9.62	9.19	-.42	4.92
	30	20	8.60	8.91	.31	6.01
	30	10	8.83	9.16	.33	5.06
Random Selection (Fuzzy cut-off)	20	20	8.72	8.92	.19	3.75
	20	10	9.25	8.97	-.28	3.96
	30	20	8.88	9.01	.12	4.13
	30	10	8.81	9.01	.20	4.09
Teacher Selection (Fuzzy cut-off)	20	20	8.16	8.96	.80	3.81
	20	10	8.82	9.02	.20	4.23
	30	20	8.20	9.06	.87	3.99
	30	10	8.22	9.14	.91	3.98

Table 4

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category III Data Sets ($r_{xx} = .69$, $r_{y,z} = .75$, N = 100)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.42	8.72	-.70	5.45
	20	10	9.82	9.12	-.70	5.45
	30	20	9.13	9.18	.05	5.02
	30	10	9.11	9.16	.05	5.02
Leave-out (Fuzzy cut-off)	20	20	9.91	8.86	-1.05	6.92
	20	10	9.72	9.19	-.53	6.01
	30	20	9.68	8.91	-.77	6.71
	30	10	9.55	9.16	-.40	6.01
Random Selection (Fuzzy cut-off)	20	20	9.32	8.92	-.40	4.45
	20	10	9.15	8.97	-.18	5.34
	30	20	9.39	9.01	-.38	4.24
	30	10	9.39	9.01	-.38	5.37
Teacher Selection (Fuzzy cut-off)	20	20	6.04	8.96	2.93	4.52
	20	10	7.72	9.02	1.29	4.92
	30	20	5.70	9.06	3.36	4.67
	30	10	7.76	9.14	1.48	5.06

Table 5

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category IV Data Sets (r_{xx} = .69, $r_{y,z}$ = .50, N = 100)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.56	8.72	-.84	5.16
	20	10	9.97	9.12	-.84	5.16
	30	20	9.66	9.18	-.47	5.03
	30	10	9.63	9.16	-.47	5.03
Leave-out (Fuzzy cut-off)	20	20	10.28	8.86	-1.41	6.69
	20	10	10.15	9.19	-.96	5.79
	30	20	9.74	8.91	-.84	7.23
	30	10	9.63	9.16	-.47	6.40
Random Selection (Fuzzy cut-off)	20	20	9.47	8.92	-.55	4.39
	20	10	9.73	8.97	-.76	4.48
	30	20	9.67	9.01	-.66	4.75
	30	10	9.51	9.01	-.50	5.56
Teacher Selection (Fuzzy cut-off)	20	20	8.04	8.96	.92	4.81
	20	10	8.88	9.02	.14	4.85
	30	20	7.60	9.06	1.47	5.39
	30	10	8.26	9.14	.88	5.72

Table 6

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category V Data Sets (r_{xx} = .84, $r_{y,z}$ = .75, N = 200)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.91	8.89	-1.02	2.58
	20	10	10.22	9.19	-1.02	2.58
	30	20	10.05	9.24	-.81	2.80
	30	10	9.88	9.07	-.81	2.80
Leave-out (Fuzzy cut-off)	20	20	10.59	9.01	-1.58	3.71
	20	10	10.05	8.98	-1.07	3.17
	30	20	10.18	9.11	-1.07	3.69
	30	10	9.76	9.02	-.74	3.30
Random Selection (Fuzzy cut-off)	20	20	9.91	8.98	-.93	2.41
	20	10	9.78	9.03	-.75	2.84
	30	20	9.69	8.85	-.84	2.38
	30	10	9.53	8.92	-.60	2.72
Teacher Selection (Fuzzy cut-off)	20	20	7.74	8.81	1.08	2.68
	20	10	8.89	9.00	.10	2.78
	30	20	7.53	8.94	1.41	2.64
	30	10	8.60	9.03	.42	2.79

Table 7

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category VI Data Sets ($r_{xx} = .84$, $r_{y|z} = .50$, N = 200)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.64	8.89	-.75	2.59
	20	10	9.95	9.19	-.75	2.59
	30	20	9.58	9.24	-.34	2.75
	30	10	9.41	9.07	-.34	2.75
Leave-out (Fuzzy cut-off)	20	20	10.30	9.01	-1.30	4.01
	20	10	9.81	8.98	-.83	3.12
	30	20	9.58	9.11	-.47	3.50
	30	10	9.45	9.02	-.43	3.21
Random Selection (Fuzzy cut-off)	20	20	9.61	8.98	-.63	2.96
	20	10	9.59	9.03	-.56	2.64
	30	20	9.37	8.85	-.51	2.74
	30	10	9.52	8.92	-.59	2.65
Teacher Selection (Fuzzy cut-off)	20	20	8.23	8.81	.58	2.57
	20	10	9.05	9.00	-.05	2.67
	30	20	8.24	8.94	.70	2.55
	30	10	8.85	9.03	.18	2.79

Table 8

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category VII Data Sets (r_{xx} = .69, $r_{y,z}$ = .75, N = 200)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.88	8.89	-.99	3.48
	20	10	10.19	9.19	-.99	3.48
	30	20	9.58	9.24	-.35	3.57
	30	10	9.42	9.07	-.35	3.57
Leave-out (Fuzzy cut-off)	20	20	10.18	9.01	-1.17	5.04
	20	10	10.09	8.98	-1.11	4.30
	30	20	9.62	9.11	-.51	4.92
	30	10	9.58	9.02	-.56	4.18
Random Selection (Fuzzy cut-off)	20	20	9.59	8.98	-.61	3.30
	20	10	9.89	9.03	-.86	3.52
	30	20	9.53	8.85	-.68	3.28
	30	10	9.59	8.92	-.67	3.53
Teacher Selection (Fuzzy cut-off)	20	20	5.91	8.81	2.90	3.31
	20	10	8.12	9.00	.87	3.58
	30	20	5.72	8.94	3.23	3.53
	30	10	7.46	9.03	1.56	3.63

Table 9

Differences Between Actual and Estimated Gains

by Data Analysis Situation and Cut-off Classification

for Category VIII Data Sets (r_{xx} = .69, $r_{y;z}$ = .50, N = 200)

Situation	Classification: Cut-off Point (%ile)	Classification: Cut-off Band (%)	Estimated Gain	Actual Gain	Difference: Mean	Difference: S.D.
Strict Cut-off	20	20	9.16	8.89	-.27	3.47
	20	10	9.47	9.19	-.27	3.47
	30	20	9.47	9.24	-.24	3.26
	30	10	9.31	9.07	-.24	3.26
Leave-out (Fuzzy cut-off)	20	20	9.75	9.01	-.74	4.63
	20	10	9.58	8.98	-.61	3.99
	30	20	9.15	9.11	-.04	4.16
	30	10	9.18	9.02	-.16	3.48
Random Selection (Fuzzy cut-off)	20	20	9.43	8.98	-.45	3.36
	20	10	9.45	9.03	-.42	3.38
	30	20	9.06	8.85	-.20	3.39
	30	10	9.27	8.92	-.34	3.16
Teacher Selection (Fuzzy cut-off)	20	20	7.29	8.81	1.52	3.13
	20	10	8.39	9.00	.61	3.41
	30	20	6.99	8.94	1.95	3.07
	30	10	8.28	9.03	.74	3.02

Footnotes for Appendices A-H

1. The notations in the Appendices are interpreted as follows:

   r_{xx} = data reliability

   r_{y1z} = correlation between pretest and teacher ratings

   r_{y1y2} = correlation between pretest and posttest

   $\bar{Y}_1$ = pretest mean

   Sy_1 = pretest standard deviation

   $\bar{Y}_2$ = posttest mean

   Sy_2 = posttest standard deviation

   $\bar{Z}$ = mean of teacher ratings

   Sz = standard deviation of teacher ratings

   $\bar{G}$ = growth mean

   Sg = growth standard deviation

2. Each data category consists of 100 data sets. For Categories I-IV, each of the 100 data sets consists of 100 simulated cases. For Categories V-VIII, each of the 100 data sets consists of 200 simulated cases.

3. S.D. in the last column refers to standard deviations for the 100 simulated data sets.

Appendix A

Characteristics of Data Sets in Category I

Characteristics	Mean	S.D.
r_{xx}	.84	--
r_{y1z}	.76	.04
r_{y1y2}	.76	.04
$\overline{Y}_1$	52.08	1.93
Sy_1	19.37	1.30
$\overline{Y}_2$	56.13	1.89
Sy_2	18.51	1.25
$\overline{Z}$	52.64	1.81
Sz	18.95	1.32
$\overline{G}$	12.96	.91
Sg	7.88	.59

Appendix B

Characteristics of Data Sets in Category II

Characteristics	Mean	S.D.
r_{xx}	.84	—
r_{y1z}	.51	.08
r_{y1y2}	.75	.04
$\overline{Y}_1$	52.31	2.00
Sy_1	19.25	1.33
$\overline{Y}_2$	56.23	1.74
Sy_2	18.47	1.24
$\overline{Z}$	52.55	1.79
Sz	18.99	1.47
$\overline{G}$	12.97	.86
Sg	7 88	.56

Appendix C

Characteristics of Data Sets in Category III

Characteristics	Mean	S.D.
r_{xx}	.69	—
r_{y1z}	.74	.05
r_{y1y2}	.59	.06
$\overline{Y}_1$	52.95	1.91
Sy_1	19.07	1.19
$\overline{Y}_2$	56.29	1.96
Sy_2	18.02	1.27
$\overline{Z}$	52.39	2.08
Sz	19.01	1.29
$\overline{G}$	12.88	.90
Sg	7.92	.61

Appendix D

Characteristics of Data Sets in Category IV

Characteristics	Mean	S.D.
r_{xx}	.69	—
r_{y1z}	.51	.07
r_{y1y2}	.59	.06
$\bar{Y}_1$	52.99	1.99
Sy_1	18.98	1.42
$\bar{Y}_2$	56.52	1.72
Sy_2	18.18	1.29
$\bar{Z}$	53.13	1.68
Sz	19.24	1.22
$\bar{G}$	12.77	.77
Sg	7.88	.56

Appendix E

Characteristics of Data Sets in Category V

Characteristics	Mean	S.D.
r_{xx}	.84	--
r_{y1z}	.76	.03
r_{y1y2}	.76	.03
$\bar{Y}_1$	52.15	1.22
Sy_1	19.46	.99
$\bar{Y}_2$	55.92	1.17
Sy_2	18.42	.85
$\bar{Z}$	52.77	1.24
Sz	19.07	1.03
$\bar{G}$	12.80	.44
Sg	7.91	.40

Appendix F

Characteristics of Data Sets in Category VI

Characteristics	Mean	S.D.
r_{xx}	.84	--
r_{y1z}	.52	05
r_{y1y2}	.76	.03
$\bar{Y}_1$	52.09	1.29
Sy_1	19.49	.90
$\bar{Y}_2$	55.81	1.15
Sy_2	18.55	.92
$\bar{Z}$	52.86	1.38
Sz	19.25	.95
$\bar{G}$	12.76	.48
Sg	7.87	.38

Appendix G

Characteristics of Data Sets in Category VII

Characteristics	Mean	S.D.
r_{xx}	.69	--
r_{y1z}	.74	.03
r_{y1y2}	.59	.05
$\overline{Y}_1$	52.88	1.22
Sy_1	19.10	.84
$\overline{Y}_2$	56.28	1.24
Sy_2	18.13	.95
$\overline{Z}$	52.53	1.23
Sz	19.23	.81
$\overline{G}$	12.82	.51
Sg	7.91	.36

Appendix H

Characteristics of Data Sets in Category VIII

Characteristics	Mean	S.D.
r_{xx}	.69	--
r_{y1z}	.50	.05
r_{y1y2}	.59	.04
$\overline{Y}_1$	52.71	1.24
Sy_1	19.11	.94
$\overline{Y}_2$	56.49	1.43
Sy_2	18.20	.88
$\overline{Z}$	52.90	1.38
Sz	19.12	.77
$\overline{G}$	12.85	.58
Sg	7.91	.39